CE3-R3: DATA WAREHOUSING AND MINING

NOTE:

1. **Answer question 1 and any FOUR questions from 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

Time: 3 Hours **Total Marks: 100**

1. Define and explain following basic data mining tasks.

a) Classification
b) Clustering
c) Prediction
d) Link Analysis
e) Time Series Analysis
f) OLAP
g) Knowledge Discovery

(7x4)

2.

a) Explain life cycle of data warehouse development

b) Describe benefits and drawbacks of a source-driven architecture for gathering of data at a data-warehouse, as compared to a destination-driven architecture.

(9+9)

3.

a) Define **decision tree.** Write and explain decision tree development algorithm with an appropriate example.

b) Define and explain Bayesian classification scheme.

c) Detail the improvements made by either **C4.5** or **CART** in the basic decision tree algorithm.

(6+6+6)

4.

a) The three types of concept hierarchies are: schema hierarchies, set grouping hierarchies and rule-based hierarchies. Briefly define each type of hierarchy (giving suitable example).

b) Suppose that a dataware house consists of the three dimensions time, doctor and patient and the two measures count and charge where charge is the fee that a doctor charges for a patient. Draw a schema diagram for the dataware using snowflake schema.

c) How can rules be extracted from a decision tree?

(6+6+6)

5.

a) Suppose half of all the transactions in a clothes shop are for purchase of jeans, and one third of all transactions in the shop are for purchase of T-shirts. Suppose also that half of the transactions that for purchase of jeans also for purchase of T-shirts. Write down all the (nontrivial) association rules you can deduce from the above information, giving support and confidence of each rule.

b) Compare the advantages and disadvantages of (i) K-means and (ii) K-medoids for clustering. Discuss a main challenge common to both the K-means and K-medoids algorithms.

c) Why is the NBC algorithm called Naive? Explain.

(6+6+6)

6.

a) Why is tree pruning useful in tree induction? What is drawback of using a separate set of samples to evaluate pruning.

b) Compare the advantages and disadvantages of eager classification verses lazy classification. Classify the following techniques into eager and lazy classification: K nearest neighbor, decision tree, Bayesian, neural network, case based reasoning.

c) A data warehouse consists of four dimensions date, spectator, location and game and the two measures are count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults or seniors with each category having its own charge rate. Draw a star schema for the data warehouse.

(6+6+6)

7. Write short notes on:

a) Data Mining Query Language

b) Iceberg queries

c) Mining Spatial Databases

(3x6)